Fusion évidentielle de références et interrogation flexible

Evidential reference fusion and flexible querying

S. Destercke[1] F. Saïs[2] R. Thomopoulos [1,3]

[1] UMR IATE (INRA-CIRAD-UM2-Supagro), Bât. 31, 2 Place P. Viala, F-34060 Montpellier Cedex 1
[2] LRI/INRIA Saclay-Île-de-France, projet Gemo, Parc Club Orsay Université,
Bât. G, 4 rue Jacques Monod, F-91893 Orsay Cedex
[3] LIRMM (CNRS & Université Montpellier II), 161 rue Ada, F-34392 Montpellier Cedex 5

Résumé :

Il arrive souvent que des données venant de plusieurs sources (bases de données, ...) réfèrent à la même entité du monde réel. Dans de tels cas, il est nécessaire, d'une part, d'identifier ces références multiples, et d'autre part de synthétiser cette information redondante de manière à ce qu'elle soit facilement manipulable. Dans cet article, nous proposons de synthétiser l'information à partir d'une série de critères en utilisant le formalisme des fonctions de croyance (i.e., la théorie de l'évidence). L'interrogation de l'information synthétisée est ensuite réalisée au moyen d'intégrales de Choquet.

Mots-clés :

fusion de références, interrogation flexible, fonctions de croyance, théorie de l'évidence.

Abstract:

It often happens that data coming from multiple sources (databases,...) refer to the same real-world entity. In such cases, it is necessary, first to identify these multiple references, and second to synthesize this redundant information in a tractable summary. In this paper, we propose to model this synthesized information by belief functions whose shapes are based on a set of criteria, using evidence theory formalism. The query of the synthesized information is then achieved by the use of Choquet integrals.

Keywords:

Information fusion, flexible querying, belief functions, evidence theory.

1 Introduction

La détection de données redondantes et leur fusion en une représentation unique sont deux des problèmes principaux rencontrés dans les systèmes d'intégration de données. Lorsque les informations viennent de sources hétérogènes, il arrive en effet souvent que plusieurs références (i.e. descriptions de données) représentent une même entité du monde réel.

Le problème de détection de redondances [5], supposé résolu dans cet article, consiste à décider, sur la base des informations contenues dans les références, si deux références distinctes représentent la même entité du monde réel (e.g., le même tableau, le même article, le même gène). Une fois les groupes de références redondantes formés se présente le problème de fusionner les références au sein de chaque groupe en une référence unique. C'est à ce problème que nous nous intéressons ici. Le fait de fusionner les références redondantes a plusieurs intérêts : d'une part il permet de réduire le nombre total de références, ce qui permet un stockage plus facile et une interrogation plus rapide de ces dernières, d'autre part la réponse renvoyée suite à une requête d'utilisateur est plus lisible, les doublons ayant été fusionnés.

Les problèmes de réconciliation et de fusion concernant souvent de gros volumes de données, il est alors souhaitable de construire des méthodes aussi automatisées que possible. D'autre part, les valeurs prises par un même attribut au sein d'un groupe de références réconciliées sont variables (par exemple, un tableau est appelé parfois "*Mona Lisa*", parfois "*Joconde*"), contiennent des erreurs (pouvant être introduites durant le processus d'acquisition des données) et peuvent être incomplètes ou manquantes (attributs non renseignés). Il est donc nécessaire de prendre cette incertitude (variabilité et imprécision) en compte lors de la construction de la référence fusionnée. Pourtant, la plupart des systèmes de fusion pro-

posés dans la littérature demandent une intervention de l'utilisateur dans le choix des valeurs des attributs de la référence fusionnée. De plus, ces systèmes (e.g. les ETLs) renvoient des références fusionnées où chaque attribut est doté d'une valeur unique [4, 12].

Dans cet article, nous proposons de modéliser l'incertitude au moyen de fonctions de croyance [1] [8], afin de pouvoir englober variabilité, imprécision et fiabilité des informations rencontrées dans un même modèle. Nous signalerons le lien entre ce modèle et le modèle possibiliste (i.e. utilisant un formalisme flou), exploré dans de précédent travaux [7]. Enfin, nous proposons d'utiliser les intégrales de Choquet [1] pour réaliser et évaluer la pertinence des requêtes effectuées sur les références fusionnées, ce qui nous permet de relier la méthode proposée à d'autres théories de l'incertain plus générales, notamment les ensembles de probabilités. La section 2 présente les outils de base utilisés dans cet article. Le problème générique de réconciliation et de fusion, ainsi que la méthode de fusion proposée, sont présentés en section 3. Enfin, la section 4 présente brièvement la méthode d'interrogation par intégrale de Choquet.

2 Préliminaires

Nous rappelons le formalisme des fonctions de croyance [8] utilisé pour décrire l'incertitude portant sur les valeurs des attributs des références fusionnées. Nous présentons ensuite l'intégrale de Choquet [1] utilisée ici pour évaluer l'adéquation de réponses à une requête flexible (i.e. où l'utilisateur peut exprimer des préférences sur les valeurs recherchées).

2.1 Fonctions de croyance

Les fonctions de croyance, introduites par Shafer [8] et plus tard reprises par Smets [11] dans son modèle des croyances transférables, sont des outils flexibles qui permettent de modéliser l'incertitude (variabilité et imprécision) concernant la valeur d'une variable X prenant ses valeurs sur un domaine (ici fini) $\mathcal{X}$.

Pour représenter une fonction de croyance, nous utiliserons le plus souvent sa distribution de masse. Une distribution de masse est une fonction $m : \wp(\mathcal{X}) \rightarrow [0,1]$ des sous-ensembles de $\mathcal{X}$ vers l'intervalle unité $[0,1]$ et telle que $\sum_{E \subseteq \mathcal{X}} m(E) = 1$, $m(E) \geq 0$ et $m(\emptyset) = 0$. Les ensembles E ayant une masse strictement positive sont appelés ensembles focaux. A partir de cette fonction, deux fonctions d'ensembles, les mesures de crédibilité et de plausibilité, sont définies [8], pour tout $A \subseteq \mathcal{X}$:

$$Bel(A) = \sum_{E, E \subseteq A} m(E)$$

$$Pl(A) = \sum_{E, E \cap A \neq \emptyset} m(E)$$

$Bel(A)$ mesure la quantité d'information qui étaye forcément A, et $Pl(A)$ la quantité d'information qui pourrait étayer A. La masse $m(E)$ s'interprète comme la probabilité de savoir uniquement que la vraie valeur se trouve dans E et dans aucun autre ensemble (plus général ou plus spécifique). Par la suite, nous noterons $\mathcal{F}_m$ l'ensemble des éléments focaux relatifs à une distribution m. Les fonctions de plausibilité et de crédibilité peuvent également s'interpréter comme des bornes de probabilités [2], générant alors un ensemble de probabilités $\mathcal{P}_m$ tel que $\mathcal{P}_m = \{P | \forall A \subseteq \mathcal{X},\ Bel(A) \leq P(A)\}$.

2.2 Intégrale de Choquet

L'intégrale de Choquet est un opérateur d'intégration défini pour des mesures non-additives. Etant donnée une distribution de masse m et une fonction [2] $u : \mathcal{X} \rightarrow [0,1]$, leurs intégrales de Choquet par rapport à la fonction de plausibilité, que nous notons $\overline{\mathbb{E}}_m$, et à la fonction de crédibilité, que nous notons $\underline{\mathbb{E}}_m$, peuvent

1. qui sont l'outil de base de la théorie de l'évidence.

2. On se restreindra ici aux fonctions à valeurs dans l'intervalle unité, ce qui en pratique est suffisant.

respectivement s'écrire comme suit :

$$\overline{\mathbb{E}}_m(u) = \sum_{E \in \mathcal{F}_m} m(E) sup_{x \in E} u(x)$$
$$\underline{\mathbb{E}}_m(u) = \sum_{E \in \mathcal{F}_m} m(E) inf_{x \in E} u(x)$$

Ces opérateurs reviennent à calculer les bornes d'espérance supérieure et inférieure de la fonction u sur l'ensemble des éléments de $\mathcal{P}_m$, c'est-à-dire $\overline{\mathbb{E}}_m(u) = \sup_{p \in \mathcal{P}_m} \mathbb{E}_p(u)$ et $\underline{\mathbb{E}}_m(u) = \inf_{p \in \mathcal{P}_m} \mathbb{E}_p(u)$, où $\mathbb{E}_p(u)$ est l'espérance mathématique de u étant donnée la distribution de probabilité p. Si u prend ses valeurs sur $[0, 1]$, elle peut être associée à une fonction d'appartenance floue [16]. Ces intégrales sont alors équivalentes à la notion de degré de croyance en un événement flou tel que défini par Smets [9].

Dans ce qui suit, u représentera une fonction de préférence dans une requête flexible exprimée par l'utilisateur. Les intégrales de Choquet seront utilisées pour évaluer l'adéquation d'une référence fusionnée à cette requête.

3 Réconciliation, modèle d'incertitude et fusion

Nous considérons N références (i.e. des descriptions de données) $ref_1, \ldots, ref_N$ provenant de M sources $S_1, \ldots, S_M$ et décrites par un ensemble $\mathcal{A} = \{A_1, \ldots, A_P\}$ de P attributs. Nous noterons $\mathcal{V}_p$ l'ensemble des valeurs que peut prendre l'attribut A_p. Une référence ref_n peut donc être décrite par une série de valeurs, notées $\mathcal{D}(ref_n) = \{v_{n1}, \ldots, v_{nP}\}$, où v_{np} est la valeur prise par l'attribut A_p. Notons qu'il peut exister des valeurs vides, dans le cas de données manquantes.

Exemple 1. *Nous considérons deux sources de données décrivant des peintures, résumées dans le tableau 1.*

3.1 Réconciliation

La réconciliation de références consiste, via un algorithme (par exemple, la méthode N2R [6]), à identifier les paires de références redondantes (i.e. qui représentent la même entité du monde réel), pour ensuite construire à partir de ces paires, par transitivité, une relation d'équivalence. Les classes d'équivalence ainsi définies représentent des sous-groupes formant une partition de $\{ref_1, \ldots, ref_N\}$ et faisant référence à une même entité. Pour obtenir une représentation unique de cette entité, il faut donc fusionner les références d'une même classe d'équivalence en une seule référence. Dans notre exemple, les paires $\{11, 12\}$, $\{12, 21\}$,$\{12, 22\}$,$\{11, 22\}$ sont jugées redondantes. Le sous-groupe généré par ces dernières, par transitivité, est $\{11, 12, 21, 22\}$. Dans la suite, nous noterons ces L sous-groupes $Rec_1, \ldots, Rec_L$, et $\mathcal{V}_{lp}$ l'ensemble des valeurs distinctes prises par un attribut A_p au sein d'un sous-groupe Rec_l, avec $p = 1, \ldots, P$ et $l = 1, \ldots, L$.

3.2 Critères du modèle d'incertitude

Dans chaque sous-groupe de références réconciliées, afin de représenter l'incertitude sur la valeur finale que devrait prendre chaque attribut de la référence fusionnée, nous allons procéder en deux étapes :

1. L'information apportée par chaque référence n'étant pas totalement fiable (sinon toutes les références redondantes seraient identiques), nous affaiblissons cette dernière de manière automatique à partir d'une série de critères en la transformant en une fonction de croyance.
2. Les fonctions de croyance ainsi obtenues sont ensuite fusionnées pour obtenir un modèle d'incertitude concernant la valeur finale de l'attribut dans la référence fusionnée, pour finalement obtenir une seule référence synthétique.

Dans la suite, nous nous concentrons sur un sous-groupe Rec_l donné, et sur un attribut A_p

Source S_1

Ref.	Musée	Adresse	Contact	Peinture
11	Lovre	Palais Royal, Paris	info@louvre.fr	La Joconde
12	Louvre	Palais Royal, Paris	0140105317	Jconde
13	Orsay	Rive gauche de la seine, Paris	0150616742	L'européenne

Source S_2

Ref.	Musée	Adresse	Contact	Peinture
21	Louvre	Rue rivoli, 75001 Paris	info@louvre.fr 0140105317	Mona Lisa
22	Le Louvre	99 Rue rivoli, Paris		La Joconde

Tableau 1 – Exemple de source de données

fixé. Soit v la valeur prise par la référence considérée au sein du sous-groupe Rec_l. Les critères suivants seront pris en compte dans la méthode proposée :

– **Homogénéité (Hom)** : l'homogénéité mesure la fréquence d'apparition de la valeur v considérée au sein du sous-groupe de références réconciliées $ref_n \in Rec_l$. Elle se calcule comme suit :

$$Hom(v) = \frac{|\{v_{np} = v | ref_n \in Rec_l\}|}{|Rec_l|}.$$

– **Similarité syntaxique (Sim)** : Nous notons $Sim(v, v')$ la mesure de similarité entre deux valeurs prises par l'attribut A_p dans les références réconciliées. Il existe plusieurs mesures de similarité dans la littérature [14], et le choix de l'une d'elles dépend souvent du contexte (données numériques ou non, structurées ou non, ...). Nous noterons $\mathcal{V}_{lp}^v = \{v^{1,v}, \ldots, v^{|\mathcal{V}_{lp}|,v}\}$ l'ensemble ordonné tel que $i < j \Rightarrow Sim(v, v^{i,v}) < Sim(v, v^{j,v})$, où les valeurs sont rangées par ordre de similarité syntaxique décroissante avec v. Notons que $v^{1,v} = v$.

– **Fiabilité des sources (α_m)** : Nous notons α_m la fiabilité de la source S_m. Elle peut être calculée, par exemple, en fonction de la dernière date de mise à jour de S_m [7].

– **Fréquence globale d'occurence (f)** : La fréquence globale d'occurence mesure la fréquence à laquelle apparaît la valeur v au sein de toutes les références. Une valeur apparaissant de nombreuses fois a en effet moins de chances de contenir une erreur typographique. Elle se calcule comme

$$f(v) = \frac{|\{v_{np} = v | n = 1, \ldots, N\}|}{N}$$

Les trois derniers critères sont pris en compte dans le modèle d'incertitude, et le premier critère dans la procédure de fusion.

3.3 Modèle d'incertitude

Etant donnée une référence $ref_n \in Rec_l$ fournissant pour l'attribut A_p une valeur $v \in \mathcal{V}_{lp}$, nous définissons d'abord un modèle d'incertitude qui prenne en compte la fréquence globale f et la similarité syntaxique Sim. Le modèle d'incertitude est construit de la manière suivante :

$$m_{ref_n,p}(\{v\}) = \frac{f(v)}{\sum_{v \in \mathcal{V}_{lp}} f(v)},$$

la normalisation évitant que la masse $m_{ref_n,p}(\{v\})$ devienne de plus en plus petite à mesure que le nombre de références grandit. Le reste de la masse est ensuite affecté à des ensembles emboîtés construits à partir de l'ensemble ordonné $\mathcal{V}_{lp}^v$, de la manière suivante :

$$m_{ref_n,p}(\{v^{1,v}, \ldots, v^{k,v}\}) = (1 - m_{ref_n,p}(\{v\}))NSim(k)$$

pour $k = 2, \ldots, |\mathcal{V}_{lp}|$ et

$$NSim(k) = \frac{Sim(v, v^{k,v})}{\sum_{v \in \mathcal{V}_{lp} \setminus v^{1,v}} Sim(v, v^{k,v})}.$$

Notons que la fonction $NSim$ est décroissante, du fait de l'ordonnancement de $\mathcal{V}_{lp}^v$.

Remarque 1. *Le modèle d'incertitude proposé suppose que l'ensemble $\mathcal{V}_{lp}^{v}$ soit ordonné. Si l'on suppose seulement $i < j \Rightarrow Sim(v, v^{i,v}) \leq Sim(v, v^{j,v})$ (i.e. en cas de préordre), chaque ensemble emboîté doit être augmenté, par rapport au précédent, non pas d'une valeur $v^{k,v}$ mais d'une classe d'équivalence contenant l'ensemble des valeurs ayant la même similarité avec v.*

Soit S_m la source dont vient la référence ref_n, et α_m la fiabilité de cette source. La prise en compte de cette fiabilité se fait simplement par une opération d'affaiblissement, qui consiste à transformer m_{ref} en m'_{ref} telle que

$$m'_{ref_n,p}(\{v^{1,v}, \ldots, v^{k,v}\}) = \alpha_m m_{ref_n,p}(\{v^{1,v}, \ldots, v^{k,v}\})$$

pour $k = 1, \ldots, |\mathcal{V}_{lp}| - 1$ et

$$m'_{ref_n,p}(\mathcal{V}_{lp}) = (1 - \alpha_m) + \alpha_m m_{ref_n,p}(\mathcal{V}_{lp})$$

ce qui revient à attribuer une probabilité $1 - \alpha_m$ au fait que la source se trompe et que la vraie valeur puisse être n'importe quel élément de $\mathcal{V}_{lp}$.

Dans le cas où une référence ref_n ne fournit aucune valeur pour l'attribut A_p (valeur manquante), le modèle d'incertitude devient

$$m_{ref_n,p}(\mathcal{V}_{lp}) = 1,$$

c'est-à-dire le modèle qui correspond à un état d'ignorance totale, qui est ensuite fusionné avec les autres. Les données manquantes sont donc considérées comme pouvant être toute valeur rencontrée dans le sous-groupe, et rendent le modèle d'incertitude final plus imprécis.

Exemple 2. *Nous considérons la référence* 11 *de l'exemple 1 et l'attribut* Peinture *(désigné par l'indice* 4*, pour faciliter la notation), en supposant* $\alpha_1 = 0.7$, $\alpha_2 = 0.9$, Sim(*Jconde,Mona Lisa*) $= 0.52$, Sim(*Jconde,La Joconde*) $= 0.81$, Sim(*La Joconde,Mona Lisa*) $= 0.38$ *(en utilisant la distance de Jaro-Winkler [15] pour mesurer la différence syntaxique). Nous avons donc* $v^{1,v} = \{$*La Joconde*$\}$ *la valeur donnée par la référence,* $v^{2,v} = \{$*Jconde*$\}$ *la valeur syntaxiquement la plus proche de* $v^{1,v}$*, et enfin* $v^{3,v} = \{$*Mona Lisa*$\}$*. De plus, nous supposons que les fréquences globales sont* $f(Jconde) = {}^{5}/_{10000}$, $f($*La Joconde*$) = {}^{30}/_{10000}$, $f($*Mona Lisa*$) = {}^{15}/_{10000}$*. Le modèle final d'incertitude pour la référence* 11 *et l'attribut* Peinture *est le suivant :*

$$m'_{ref_{11},4}(\{\textit{La Joconde}\}) = 0.7 \times \frac{7}{10},$$
$$m'_{ref_{11},4}(\{\textit{La Joconde,Jconde}\}) = 0.7 \times (\frac{3}{10}) \times \frac{0.81}{1.19},$$
$$m'_{ref_{11},4}(\mathcal{V}_{l4}) = (1 - 0.7) + 0.7 \times (\frac{3}{10}) \times \frac{0.38}{1.19},$$

avec $\mathcal{V}_{l4} = \{$*La Joconde,Jconde,Mona Lisa*$\}$.

Remarque 2. *La fonction de croyance définie dans cette section a des ensembles focaux emboîtés, et est donc formellement équivalente à la distribution de possibilité telle que* $\pi_{ref_n,p}(x) = pl_{ref_n,p}(\{x\}) \forall x \in \mathcal{V}_{lp}$*. Cette distribution accorde à la valeur v donnée par la référence le plus haut niveau de possibilité, i.e.* $\pi_m(\{v\}) = 1$*, pour ensuite attribuer des valeurs décroissantes selon l'ordre de similarité syntaxique.*
Notons que les deux modèles contiennent, à cette étape, la même information, et il est donc possible d'utiliser directement des distributions de possibilité, moins gourmandes en espace de stockage et pouvant bénéficier d'implémentations utilisées pour la logique floue, sans perte d'information.

3.4 Fusion des modèles

Nous proposons ensuite de fusionner l'ensemble des modèles issus de chaque référence du groupe Rec_l en utilisant l'opérateur de moyenne arithmétique (équipondérée). Cela revient à construire le modèle $m_{\Sigma_l,p}$ tel que, pour $E \subseteq \mathcal{V}_{lp}$,

$$m_{\Sigma_l,p}(E) = \sum_{ref \in Rec_l} \frac{1}{|Rec_l|} m_{ref}(E),$$

avec m_{ref} le modèle d'incertitude construit pour la référence ref. Cet opérateur de fusion, qui revient à pratiquer un comptage, permet d'intégrer au modèle final le critère d'homogénéité Hom, que nous avions jusqu'ici

laissé de côté. En effet, plus une même valeur d'un attribut sera donnée par les références au sein de Rec_l, plus elle sera comptabilisée dans la procédure de fusion.

Exemple 3. *Poursuivons l'exemple 2, en considérant toujours les mêmes données et l'attribut* Peinture. *Nous obtenons les modèles suivants :*

$$m'_{ref_{11},4}(\{\textit{La Joconde}\}) = 0.49,\ \ m'_{ref_{11},4}(\mathcal{V}_{l4}) = 0.37,$$
$$m'_{ref_{11},4}(\{\textit{La Joconde,Jconde}\}) = 0.14,$$
$$m'_{ref_{12},4}(\{\textit{Jconde}\}) = 0.07,\ \ m'_{ref_{12},4}(\mathcal{V}_{l4}) = 0.50,$$
$$m'_{ref_{12},4}(\{\textit{La Joconde,Jconde}\}) = 0.43,$$
$$m'_{ref_{21},4}(\{\textit{Mona Lisa}\}) = 0.27,\ \ m'_{ref_{21},4}(\mathcal{V}_{l4}) = 0.37,$$
$$m'_{ref_{21},4}(\{\textit{Mona Lisa, Jconde}\}) = 0.36,$$
$$m'_{ref_{22},4}(\{\textit{La Joconde}\}) = 0.54,\ \ m'_{ref_{22},4}(\mathcal{V}_{l4}) = 0.21,$$
$$m'_{ref_{22},4}(\{\textit{La Joconde,Jconde}\}) = 0.25,$$

avec $\mathcal{V}_{l4} = \{$*La Joconde,Jconde,Mona Lisa*$\}$. *Au final, le modèle* $m_{\Sigma_l,4}$ *est :*

$$m_{\Sigma_l,4}(\{\textit{La Joconde}\}) = 0.26,\ \ m_{\Sigma_l,4}(\{\textit{Mona Lisa}\}) = 0.07,$$
$$m_{\Sigma_l,4}(\{\textit{La Joconde,Jconde}\}) = 0.2,$$
$$m_{\Sigma_l,4}(\{\textit{Mona Lisa, Jconde}\}) = 0.09,$$
$$m_{\Sigma_l,4}(\{\textit{Jconde}\}) = 0.02,\ \ m_{\Sigma_l,4}(\mathcal{V}_{l4}) = 0.36.$$

Après fusion, nous obtenons donc L références fusionnées $ref_{\Sigma_l}, l = 1, \ldots, L$ telles qu'à chaque attribut $A_p, p = 1, \ldots, P$ de la référence ref_{Σ_l} est associée une distribution de masses $m_{\Sigma_l,p}$ qui décrit l'incertitude sur la vraie valeur que devrait prendre l'attruibut A_p.

Remarque 3. *Notons que* $pl_{\Sigma_l,p}(E) = \sum_{ref \in Rec_l} \frac{1}{|Rec_l|} pl_{ref,p}(E)$ *pour tout* $E \subseteq \mathcal{V}_{lp}$, *avec* $pl_{\Sigma_l,p}, pl_{ref,p}$ *les mesures de plausibilité respectivement induites par* $m_{\Sigma_l,p}$ *et* $m_{ref,p}$. *La moyenne arithmétique peut directement s'effectuer sur les mesures de plausibilité.*

Compte tenu de la remarque 2, le modèle possibiliste $\pi_{\Sigma_l,p}$ *tel que, pour tout* $x \in \mathcal{V}_{lp}$, $\pi_{\Sigma_l,p} = \sum_{ref \in Rec_l} \frac{1}{|Rec_l|} \pi_{ref,p}(x)$, *avec* $\pi_{ref,p}(x) = pl_{ref,p}(x)$, *coincide avec le modèle évidentiel sur les singletons, et reste donc cohérent avec ce dernier. Néanmoins, les ensembles focaux de* $m_{\Sigma_l,p}$ *ne seront plus, en général, emboîtés. Considérer uniquement* $\pi_{\Sigma_l,p}$ *reviendrait donc à ne considérer qu'une part de l'information fournie par* $m_{\Sigma_l,p}$, *i.e., celle concernant la plausibilité des singletons.*

$\pi_{\Sigma_l,p}$ *étant plus simple à manipuler et moins lourd à stocker en mémoire que* $m_{\Sigma_l,p}$, *le choix de retenir ce modèle appauvri peut se justifier par des raisons calculatoires.*

4 Requête flexible et ordonnancement des résultats

Une requête (flexible) $\mathcal{Q}$ est composée de deux éléments :

- un ensemble $\mathcal{A}^{\mathbb{P}} \subseteq \mathcal{A}$ de Pr attributs de projection dont les valeurs seront présentées à l'utilisateur ;
- un ensemble $\mathcal{A}^{\mathbb{S}} \subseteq \mathcal{A}$ de Se attributs de sélection, sur lesquels est défini un ordre (total) de préférence $\prec$ spécifié par l'utilisateur. Dans la suite, nous considérons que les attributs de sélections $A^{\mathbb{S}}_1, \ldots, A^{\mathbb{S}}_{Se}$ sont indexés de telle manière que $i < j \Rightarrow A^{\mathbb{S}}_i \prec A^{\mathbb{S}}_j$.

Soit $\mathcal{V}^{\mathbb{S}}_i$ l'ensemble des valeurs que peut prendre $A^{\mathbb{S}}_i$. Dans une requête flexible, à chaque attribut $A^{\mathbb{S}}_i \in \mathcal{A}^{\mathbb{S}}, i = 1, \ldots, Se$ est associée une fonction de préférence $\mu_{A^{\mathbb{S}}_i} : \mathcal{V}^{\mathbb{S}}_i \rightarrow [0, 1]$, formellement équivalente à un ensemble flou et traduisant quelles valeurs sont jugées les plus pertinentes par l'utilisateur. Nous noterons $\mathcal{A}_{\prec} = \{A_{(1)}, \ldots, A_{(P)}\}$ l'ensemble des attributs réordonnés [3] tel que $A_{(i)} = A^{\mathbb{S}}_i$.

Exemple 4. *Un utilisateur souhaite récupérer les musées référençant une peinture particulière. Il spécifie donc une requête* $\mathcal{Q}$ *telle que* $\mathcal{A}^{\mathbb{P}} = \{$*Musée, Peinture*$\}$, *les valeurs qu'il souhaite récupérer, et telle que* $\mathcal{A}^{\mathbb{S}} = \{Peinture\}$, *l'attribut sur lequel va porter la sélection. Nous avons donc* $A_{(1)} = \{Peinture\}$, *l'ordre sur les autres attributs pouvant être quelconque.*

L'utilisateur souhaite faire une recherche sur la Joconde, mais sait que ce tableau peut aussi être référencé comme Mona Lisa. Il fournit donc $\mu_{Peinture}($*La Joconde*$) = 1$ *et*

3. L'ordre des attributs dans $\mathcal{A} \setminus \mathcal{A}^{\mathbb{S}}$ peut être quelconque

$\mu_{Peinture}$(*Mona Lisa*) $= 0.5$ *(et* $\mu_{Peinture} = 0$ *pour toute autre valeur).*

Etant donnée une requête $\mathcal{Q}$, une première étape est de calculer l'ensemble $\mathcal{R}ef_{\mathcal{Q}} \subseteq \{ref_{\Sigma_l} | l = 1, \ldots, L\}$ des références fusionnées retenues, qui est tel que

$$\mathcal{R}ef_{\mathcal{Q}} = \{ref_{\Sigma_l} | \exists j, \overline{\mathbb{E}}_{m_{\Sigma_l,(j)}}(\mu_{A_j^{\mathbb{S}}}) > 0\}.$$

Il s'agit de toutes les références qui satisfont potentiellement les préférences exprimées par la requête de l'utilisateur, i.e., dont l'intégrale de Choquet de la fonction de plausibilité est positive.

Les références retenues sont ensuite ordonnées, selon la proposition de Dubois et Prade [3], comme suit :

1. fixer $j = 1$, calculer pour $l = 1, \ldots, L$ $\underline{\mathbb{E}}_{m_{\Sigma_l,(j)}}(\mu_{A_j^{\mathbb{S}}}), \overline{\mathbb{E}}_{m_{\Sigma_l,(j)}}(\mu_{A_j^{\mathbb{S}}})$;
2. ordonner les références ref_{Σ_i} selon les valeurs décroissantes de $\underline{\mathbb{E}}_{m_{\Sigma_l,(j)}}(\mu_{A_j^{\mathbb{S}}})$;
3. si des ex-aequo existent (l'ordre obtenu est un pré-ordre), raffiner l'ordre par valeurs décroissantes de $\overline{\mathbb{E}}_{m_{\Sigma_l,(j)}}(\mu_{A_j^{\mathbb{S}}})$;
4. s'il reste des ex-aequo, prendre $j = 2$, et recommencer depuis le début.

Notons que ce critère de décision est proche d'un critère maximin lexicographique. D'autres critères pourraient être utilisés pour ordonner les réferences retenues, par exemple la probabilité pignistique [10] apparaît comme un choix naturel ici, l'incertitude étant modélisée par des fonctions de croyance. De même, le lien entre fonctions de croyances et probabilités imprécises permet le choix de nouveaux critères [13]. Dans ce dernier cas et selon le critère de décision choisi (e.g., E-admissibilité), l'ordre induit peut être partiel, ce qui peut poser des problèmes pour ordonner les résultats. Il serait donc intéressant de mener des tests pour étudier dans quelles situations un critère est jugé le plus pertinent par les utilisateurs.

Pour simplifier la présentation des résultats, nous proposons ensuite de fournir, pour chaque référence fusionnée dans $\mathcal{R}ef_{\mathcal{Q}}$, seulement la valeur possédant le degré de plausibilité le plus élevé de chaque attribut de projection.

Exemple 5. *Considérons la requête de l'exemple 4. La seule référence fusionnée à être dans* $\mathcal{R}ef_{\mathcal{Q}}$ *est la référence de l'exemple 3, donc la réponse à la requête est :*

$$< \textit{Louvre, La Joconde} >$$

qui, étant données les préférences définies et le modèle obtenu à l'exemple 3, a comme valeurs d'intégrales de Choquet

$$\underline{\mathbb{E}}_{m_{\Sigma_l,(1)}}(\mu_{Peinture}) = 0.26 \times 1 + 0.07 \times 0.5 = 0.295$$

$$\overline{\mathbb{E}}_{m_{\Sigma_l,(1)}}(\mu_{Peinture}) = (0.26 + 0.2 + 0.3) + (0.09 + 0.07) * 0.5 = 0.84$$

5 Conclusion

Nous avons présenté dans cet article une méthode de fusion de références réconciliées fondée sur la théorie des fonctions de croyance (ou théorie de l'évidence). Le choix des fonctions de croyance comme modèle permet de prendre aisément en compte la variabilité des données, leurs fiabilité incomplète et l'absence de données (via le modèle d'ignorance). Les fonctions de croyance apparaissent comme un bon compromis entre facilité d'utilisation et flexibilité du modèle. Dans le cas où le temps de calcul doit être réduit, il est facile de se ramener à un modèle possibiliste, moins riche mais plus facile à manipuler. Il est également possible de rapprocher la méthode proposée de modèles s'appuyant sur des ensembles de probabilités, lui donnant un caractère plus générique.

En termes de perspectives à ce travail, nous pouvons citer :

– l'implémentation et le test de la proposition sur des bases de données réelles ;
– la prise en compte de connaissances supplémentaires sur le domaine d'application, par exemple via des ontologies ;

– la comparaison de différents critères d'ordonnancement des réponses aux requêtes, en particulier leur pertinence jugée par l'utilisateur.

Enfin nous envisageons d'étendre la méthode proposée à la prise en compte de données multivaluées. Par exemple, la valeur de l'attribut "Contact" dans la référence 21 fournit un cas de donnée multivaluée, comportant à la fois une adresse électronique et un numéro de téléphone. Dans le cas de données multivaluées, plusieurs valeurs parmi celles figurant dans les données peuvent être les bonnes, et la référence fusionnée ne fournit plus un ensemble de valeurs exclusives.

Références

[1] G. Choquet. Theory of capacities. *Annales de l'institut Fourier*, 5 :131–295, 1954.

[2] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38 :325–339, 1967.

[3] D. Dubois and H. Prade. *Fuzziness in Database Management Systems*, chapter Tolerant fuzzy pattern matching : an introduction, pages 42–58. Physica-Verlag, 1995.

[4] Y. Papakonstantinou, S. Abiteboul, and H. Garcia-Molina. Object fusion in mediator systems. In *VLDB*, pages 413–424, San Francisco, CA, USA, 1996.

[5] F. Saïs, N. Pernelle, and M.-C. Rousset. L2r : A logical method for reference reconciliation. In *AAAI*, pages 329–334, 2007.

[6] F. Sais, N. Pernelle, and M.-C. Rousset. Combining a logical and a numerical method for data reconciliation. *Journal of Data Semantics*, 12 :66–94, 2009.

[7] F. Saïs and R. Thomopoulos. Reference fusion and flexible querying. In *ODBASE-OTM Conferences (2)*, pages 1541–1549, 2008.

[8] G. Shafer. *A mathematical Theory of Evidence*. Princeton University Press, New Jersey, 1976.

[9] P. Smets. The degree of belief in a fuzzy event. *Information Science*, 25 :1–19, 1981.

[10] P. Smets. Decision making in the tbm : the necessity of the pignistic transformation. *I.J. of Approximate Reasoning*, 38 :133–147, 2005.

[11] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66 :191–234, 1994.

[12] V.S. Subrahmanian, S. Adali, A. Brink, R. Emery, J. L. Lu, A. Rajput, T. J. Rogers, R. Ross, and C. Ward. Hermes : A heterogeneous reasoning and mediator system, 1995.

[13] M.C.M. Troffaes. Decision making under uncertainty using imprecise probabilities. *Int. J. of Approximate Reasoning*, 45 :17–29, 2007.

[14] P. Ravikumar W. Cohen and S.E. Fienberg. A comparison of string metrics for matching names and records. In *Proc. of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, 2003.

[15] W.E. Winkler. The state of record linkage and current research problems. Technical report, Statistics of Income Division, Internal Revenue Service Publication R99/04, 1999.

[16] L.A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning-i. *Information Sciences*, 8 :199–249, 1975.